

On the monotonization of the training set

Rustem S. Takhanov

Abstract

We consider the problem of minimal correction of the training set to make it consistent with monotonic constraints. This problem arises during analysis of data sets via techniques that require monotone data. We show that this problem is NP-hard in general and is equivalent to finding a maximal independent set in special orgraphs. Practically important cases of that problem considered in detail. These are the cases when a partial order given on the replies set is a total order or has a dimension 2. We show that the second case can be reduced to maximization of a quadratic convex function on a convex set. For this case we construct an approximate polynomial algorithm based on convex optimization.

Keywords: machine learning, supervised learning, monotonic constraints.

1 Introduction

Requirements to a classifying rule in supervised learning problems consist of two parts. The first part is induced by a set of precedents, called the training set. Each element in the training set is a pair of "object-reply" type. A classifying rule which is a mapping from objects set to the replies set should map objects from the training set pairs to the consistent replies. And the second part of requirements express our common knowledge of a classifying rule. One of the popular types of such requirements is the monotonicity which is considered in that paper. In some cases these two parts of requirements can not be satisfied both and then we have a problem of a minimal correction of the training set. Let us see what that problem is.

Suppose the sets X, Y are given and on this sets we have partial orders \geq^X, \geq^Y consistently. We assume more that the partial order \geq^Y is a lattice. For any given mapping $o : X' \rightarrow Y$ where $X' \subseteq X, |X'| < \infty$ we pose a problem of finding a function $f : X \rightarrow Y$ which is monotone due to partial orders \geq^X, \geq^Y and minimizes the following functional: $Er_o(f) = |\{x | f(x) \neq o(x)\} \cap X'|$.

Let us denote the set of monotonic functions from X to Y by $M(\geq^X, \geq^Y)$. Then for a given mapping $o : X' \rightarrow Y$ our task is the following:

$$Er_o(f) \rightarrow \min_{f \in M(\geq^X, \geq^Y)}$$

Every mapping $f' : X' \rightarrow Y$ which is monotone on the subset $X' \subseteq X$ can be extended to the mapping monotone on the whole set X because (Y, \geq^Y) is a lattice. Actually on every finite subset of the lattice (Y, \geq^Y) the operation sup is defined and the function $f(x) = \sup \{f'(x') | x' \in X', x' \leq^X x\}$ is both monotone and satisfies $f(x) = f'(x), x \in X'$. From this we see that in the posed problem we can imply that $X' = X$. From the above said we conclude more that this problem is equivalent to finding a maximal subset $X'' \subseteq X'$ such that the function o restricted on the subset X'' is monotone.

So let us consider the following generalization of our problem which we will call MaxCMS(Maximal Consistent with Monotonicity Set).

MaxCMS. The finite sets B_n, B_m where $B_r = \{1, \dots, r\}$ are given; on each of them partial orders \geq^1, \geq^2 are defined consistently and the function $\varphi : B_n \rightarrow B_m$ is given. Then every element $i \in B_n$ is assigned by a positive integer weight w_i . Our task is to find a maximal by weight subset $B \subseteq B_n$ such that the function φ restricted on B is monotone i.e. $\forall i, j \in B [i \geq^1 j \rightarrow \varphi(i) \geq^2 \varphi(j)]$.

Definition 1. The set $B \subseteq B_n$ is called acceptable iff the function φ restricted on B is monotone.

Definition 2. A set which is acceptable and maximal by weight is denoted by $MaxCMS(\geq^1, \geq^2, \varphi, w)$ (in some cases we use this notation to mean the weight of this set).

In the remainder of the paper we will consider that problem.

2 Training set monotonization and maximal independent sets

In this section we will show that MaxCMS is equivalent to finding a maximal independent set (or minimal vertex cover) in special orgraphs.

Definition 3. Let $G = (V, E)$ be an orgraph and every vertex v of an orgraph has a positive integer weight w_v . A set of vertexes is called independent iff every pair of its elements is not connected by an edge. The maximal by weight independent set is denoted by $IS(G, w)$ (in some cases we use this notation to mean the weight of this set).

As well-known, the supplement of independent set is vertex cover.

Let us define the following partial preorder on B_n (recall, that it means transitive and reflexive binary predicate):

$$i \succ j \Leftrightarrow \varphi(i) \geq^2 \varphi(j).$$

Consider the orgraph $G = (V, E)$ with $V = B_n$ and $E = \{(i, j) \mid i \geq^1 j, \varphi(i) \not\geq^2 \varphi(j)\}$. The orgraph G can alternatively be defined through the following equalities: $V = B_n$ and $E = \geq^1 \cap \overline{\succ}$ where $\overline{\succ}$ is a supplement of the binary predicate.

Definition 4. An orgraph which has the edge set represented as a intersection of a partial order and a supplement of a partial preorder is called special.

Theorem 1. The maximal acceptable set is equal to the maximal independent set of the special orgraph G , i.e. $MaxCMS(\geq^1, \geq^2, \varphi, w) = IS(G, w)$.

Proof. Any independent set B of the orgraph G satisfies the condition: if $i, j \in B$ and $i \geq^1 j$ then $\varphi(i) \geq^2 \varphi(j)$, i.e. the function φ restricted on B is monotone. The inverted statement is correct also: if restriction of φ on B is monotone then B is an independent set in G . From this we obtain the proposition of the theorem.

Theorem 2. Let the special orgraph G' be defined by the vertex set $V' = B_n$ with weights w'_i and the edge set $E' = \geq' \cap \overline{\succ'}$; both \geq' and $\overline{\succ'}$ are given (i.e. the edge set E' need not be decomposed). Then the problem of finding maximal independent set in such an orgraph polynomially reducible to MaxCMS.

Proof. Let us divide the set V' on the equivalence classes due to predicate $x \sim y \Leftrightarrow x \succ' y \& y \succ' x$. Then we can naturally define the corresponding mapping $\varphi' : V \rightarrow V / \sim$. On the factor-set V / \sim it is induced the partial order $\bar{x} \geq'' \bar{y} \Leftrightarrow x \succ' y$. It is easy to see that $IS(G', w') = MaxCMS(\geq', \geq'', \varphi', w')$. The reduction is done in $O(n^2)$ steps.

3 NP-hardness of MaxCMS.

In the previous chapter it was shown that MaxCMS is equivalent to finding maximal independent set(or minimal vertex cover) in special orgraphs. The problem of finding an acceptable set of cardinality more than C is denoted by CMS. Obviously, it is in NP.

Theorem 3. CMS is NP-complete.

Proof. Let us reduce CMS to 3-SAT using the trick from [2].

Let 3-CNF be given with $U = \{u_1, \dots, u_n\}$ being the set of variables used in it. Let $C = \{c_1, \dots, c_m\}$ be the set of clauses such that each clause consists of 3 literals that differ by their variables (literal is symbol u_i or $\overline{u_i}$). For every clause we order literals that belong to it. Then the fact of meeting the literal l on the s -th place in the clause c_r is denoted by lc_r^s . Let us consider the orgraph such that its vertex set is a union of all literals and threefold copies of clauses $V = \{u_1, \overline{u_1}, \dots, u_n, \overline{u_n}\} \cup \{c_1^1, c_1^2, c_1^3, \dots, c_m^1, c_m^2, c_m^3\}$. Let us define the edge set being equal to $E = E_1 \cup E_2$ where $E_1 = \{(u_i, \overline{u_i})\}_{i=1}^n \cup \{(u_k, c_m^l) | u_k c_m^l\} \cup \{(c_m^l, \overline{u_k}) | \overline{u_k} c_m^l\}$ and $E_2 = \{(c_j^1, c_j^2), (c_j^2, c_j^3), (c_j^1, c_j^3)\}_{j=1}^m$ (later we will need this division of the edge set on 2 subsets).

A vertex cover of the orgraph $G = (V, E)$ of the cardinality $n + 2m$ exists iff the original 3-CNF is satisfiable. Actually, one from every pair of vertexes $u_i, \overline{u_i}$ and two from every triple c_j^1, c_j^2, c_j^3 should fall into the vertex cover, because they are pairwise connected. And so, the cardinality of a vertex cover is not less than $n + 2m$.

Suppose the vertex cover of the required cardinality exists. If the literal u_i is in it we define $u_i = \text{true}$, otherwise $u_i = \text{false}$. All variables should be initialized in this manner, because from the above said it is clear that u_i or $\overline{u_i}$ is in the cover excluding both of them. Then this assignment, as easily seen, satisfies the original 3-CNF. This reasoning can be inverted and we obtain that the existence a satisfying assignment is equivalent to the existence of a vertex cover of the cardinality $n + 2m$.

Now let us consider the orgraph $G' = (V, E^* \setminus E)$ where E^* is a transitive closure of E . Suppose that the edge set of G' is transitive. Then defining $\geq = E^*$ and $\succ = E^* \setminus E$ we obtain that $\geq \cap \succ = E$. This means that our problem is reduced to finding the minimal vertex cover, and consequently, the maximal independent set of the special orgraph $G = (V, E)$, which is by theorem 2 is equivalent to MaxCMS, or CMS when $C = 2n + 3m - (n + 2m) = n + m$.

Let us show that the edge set of G' is transitive. As E^* is transitive, $E^* \setminus E$ is not transitive only if there exists such $(u, v), (v, t) \in E^* \setminus E$ that $(u, t) \in E$. Let $(u, t) \in \{(u_i, \overline{u_i})\}_{i=1}^n$. It is easy to see that any path in the G starting with u_i can not end with literal $\overline{u_i}$, because otherwise there should exist a clause that contains both u_i and $\overline{u_i}$. Let us now consider the case when $(u, t) \in \{(c_j^1, c_j^2), (c_j^2, c_j^3), (c_j^1, c_j^3)\}$. In that case the path starting from c_j^α and finishing in c_j^β can not contain an element which does not belong to $\{c_j^1, c_j^2, c_j^3\}$. Consequently, $(u, v), (v, t) \in E$, which contradicts to $(u, v), (v, t) \in E^* \setminus E$. And the last case is when $(u, t) \in \{(u_k, c_m^l) | u_k c_m^l\}$. But every path in orgraph G which starts in u and finishes in t is equal to edge (u, t) , and this means $(u, v) \notin E^* \setminus E$. In the same manner the case $(u, t) \in \{(c_m^l, \overline{u_k}) | \overline{u_k} c_m^l\}$ is considered. So, the set $E^* \setminus E$ is transitive and the reduction of 3-CNF to CMS is done.

4 1-MaxCMS

Any partial order on a finite set can be represented as intersection of total orders.

Definition. Let the partial order \geq be given on the set M . The minimal number d such that \geq is an intersection of total orders \geq_1, \dots, \geq_d , i.e. $\geq = \geq_1 \cap \dots \cap \geq_d$, is called the dimension of \geq .

Consider MaxCMS with input $(\geq^1, \geq^2, \varphi, w)$ in case when the dimension of \geq^2 is equal to d . In that case $\geq^2 = \geq_1 \cap \dots \cap \geq_d$. The consistent special orgraph $G = (V, E)$ satisfies: $V = B_n$ and $E = \geq^1 \cap \overline{\succ}$ where $i \succ j \Leftrightarrow i \succ_1 j \& \dots \& i \succ_d j$ and $i \succ_s j \Leftrightarrow \varphi(i) \geq_s \varphi(j)$. And then,

$$E = \geq^1 \cap \overline{\succ_1 \cap \dots \cap \succ_d} = \geq^1 \cap (\overline{\succ_1} \cup \dots \cup \overline{\succ_d}) = (\geq^1 \cap \overline{\succ_1}) \cup \dots \cup (\geq^1 \cap \overline{\succ_d}).$$

As each predicate $\overline{\succ_s}$ is transitive, E is a union of d transitive predicates.

Definition. The problem MaxCMS with input $(\geq^1, \geq^2, \varphi, w)$ for case when the dimension of \geq^2 is equal to d is called d -MaxCMS.

In fact, the above mentioned showed that

Theorem 4. The problem d -MaxCMS is reduced to finding the maximal independent set in the orgraph $G = (V, E)$ where $E = \succ^1 \cup \dots \cup \succ^d$ and predicates \succ^s are transitive and there are no cycles in G .

From the theorem 4 we see that 1-MaxCMS is reduced to finding the maximal independent set in the circuit-free orgraph $G = (V, E)$ that has the edge set satisfying the following transitivity rule: if $(u, v), (v, t) \in E$ then $(u, t) \in E$. This problem is polynomially tractable because the graph that can be obtained from G by transformation of oriented edges to non-oriented is a comparability graph of some partial order which is known to be perfect. We will adduce one of the proofs of the tractability due to [4].

Theorem 5. 1-MaxCMS is polynomially tractable.

Proof. Defining $x \triangleright y \Leftrightarrow (x, y) \in E$, the orgraph can be seen as partially ordered set (V, \triangleright) . The algorithm solves the problem via reducing it to the task of minimizing a flow in some circuit-free network. Let us denote the sets of minimal and maximal elements of (V, \triangleright) by $\min G$ and $\max G$ consistently. For every vertex $v \in V$ of the orgraph G we introduce 2 copies v^+, v^- . And then we define $V' = \{v^+, v^-\}_{v \in V} \cup \{s, t\}$ and $E' = \{(v^+, v^-)\}_{v \in V} \cup \{(x^-, y^+) \mid (x, y) \in E\} \cup \{(s, a^+) \mid a \in \min G\} \cup \{(b^-, t) \mid b \in \max G\}$. We obtained the orgraph $G' = (V', E')$. The minimal flow through the edge (v^+, v^-) is defined to be equal to the corresponding weights w_v , and for other edges it equals 0. The maximal flow through every edge is ∞ . It is easy to see that for every edge $e \in E'$ of the orgraph G' we can find a path from s to t that goes through e . It is well-known that under that condition we can apply the min flow-max cut theorem.

The minimal flow of given network, that can be obtained via modified Ford-Fulkerson algorithm (common algorithm finds maximal flow), corresponds to the maximal W-cut (common algorithm finds minimal cut), where by the weight of a cut we mean the following expression:

$$\sum_{(u,v) \in E, u \in S, v \in \overline{S}} c_{\min}(e) - \sum_{(u,v) \in E, v \in S, u \in \overline{S}} c_{\max}(e).$$

Note that the weight of a cut is defined differently from the sum of weights between parts of a cut and that is why we call the problem maximal W-cut. Consider any cut $V' = S \cup \overline{S}$ where $s \in S, t \in \overline{S}$ with the weight different from $-\infty$. Since maximal flow

through edges is ∞ , for every edge $(u, v) \in E'$ there can not be $v \in S, u \in \overline{S}$. And edges $(u, v) \in E'$ for $u \in S, v \in \overline{S}$ can make a contribution to the weight of a cut only when $u = r^+, v = r^-$. Let us denote $R = \{r | r^+ \in S, r^- \in \overline{S}\}$. Obviously, the elements of R constitute an independent set in G and the weight of a cut is exactly equal to the weight of the set. The conversion of the statement is also correct, i.e. every independent set R of G correspond to the cut $S = \{u^+, u^- | u \notin R \& \exists r \in R [r \triangleright u]\} \cup \{r^+ | r \in R\} \cup \{s\}$, the weight of a cut being equal to the weight of R . From this it is clear that the result of an algorithm will be the maximal cut that correspond to the maximal independent set in G . The theorem proved.

The task of finding the minimal flow can be written in the LP form:

$$\begin{aligned} x(\Gamma) &\geq 0, \Gamma \in G(s, t) \\ \sum_{\Gamma \in G(v)} x(\Gamma) &\geq w_v \\ \sum_{\Gamma \in G(s, t)} x(\Gamma) &\rightarrow \min \end{aligned}$$

where $G(s, t)$ is a set of paths in orgraph $G' = (V', E')$ from s to t , and $G(v) \subset G(s, t)$ is a set of paths going through the edge (v^+, v^-) . In the dual form:

$$\begin{aligned} y(v) &\geq 0, v \in V \\ \sum_{(v^+, v^-) \in \Gamma} y(v) &\leq 1, \Gamma \in G(s, t) \\ \sum_{v \in V} w_v y(v) &\rightarrow \max \end{aligned}$$

From the above stated we conclude that the dual problem always has a boolean solution. Polyhedron of the dual problem is denoted by $\Pi(G)$.

5 2-MaxCMS

Now we will consider the problem 2-MaxCMS. This problem arise when a partial order on the replies set is not total, but, for example, has a tree structure. As we know, it can be reduced to finding the maximal independent set in the circuit-free orgraph $G = (V, E)$ where $E = \succ^1 \cup \succ^2$ and the predicates \succ^s are transitive. From now on we will consider just that problem.

Note that edges of the circuit-free orgraph from theorem 3 are also divided on 2 sets E_1 and E_2 , both of them being transitive. From this we conclude that the problem is NP-hard.

Consider 2 orgraphs: $G_1 = (V, \succ^1)$ and $G_2 = (V, \succ^2)$. Note that the maximal independent set of the orgraph $G = (V, E)$ is also an independent set in both G_1 and G_2 . Then the following theorem is obvious.

Theorem 6. The set of solutions to the following quadratic programming problem

$$\begin{aligned} \bar{x} &\in \Pi(G_1) \\ \bar{y} &\in \Pi(G_2) \\ \psi(\bar{x}, \bar{y}) &= \sum_{v \in V} w_v x_v y_v \rightarrow \max \end{aligned}$$

contains such boolean \bar{x}^*, \bar{y}^* that $\{v | x_v^* y_v^* = 1\}$ is the maximal independent set in G .

Proof. With fixed \bar{x} (fixed \bar{y}) the maximum of $\sum_{v \in V} w_v x_v y_v$ is reached on some boolean \bar{y} (boolean \bar{x}). It means that the maximum by both vectors can be achieved with boolean values of components.

Theorem 7. The following is true

$$\max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \psi(\bar{x}, \bar{y}) = \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \gamma(\bar{x}, \bar{y}),$$

where

$$\gamma(\bar{x}, \bar{y}) = \frac{1}{2} \sum_{v \in V} w_v (x_v + y_v)^2 - w_v (x_v + y_v)$$

Proof.

$$\begin{aligned} \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \sum_{v \in V} w_v x_v y_v &= \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \frac{1}{2} \sum_{v \in V} w_v (x_v + y_v)^2 - w_v (x_v^2 + y_v^2) \geq \\ &\geq \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \frac{1}{2} \sum_{v \in V} w_v (x_v + y_v)^2 - w_v (x_v + y_v) \end{aligned}$$

Since maximum of the left part of inequality is achieved on boolean vectors, it is clear that the equality holds. Taking into account that the functional $\gamma(\bar{x}, \bar{y})$ is convex, we see that the problem was reduced to the maximization of a convex quadratic function on a convex set.

Consider the functional

$$\varphi(\bar{x}, \bar{y}) = -\frac{1}{2} \sum_{v \in V} w_v (x_v - y_v)^2 - w_v (x_v + y_v)$$

Theorem 8. The following is true

$$\max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \varphi(\bar{x}, \bar{y}) \geq \max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \psi(\bar{x}, \bar{y}),$$

the values of $\varphi(\bar{x}, \bar{y})$ and $\psi(\bar{x}, \bar{y})$ being equal on the boolean vectors of the polyhedron $\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)$.

Proof. The verification of the second statement is obvious. The first follows it, because the maximum of the right part by theorem 6 can be achieved on boolean vectors.

Consider the following optimization task:

$$\begin{aligned} \bar{x} &\in \Pi(G_1) \\ \bar{y} &\in \Pi(G_2) \\ \varphi(\bar{x}, \bar{y}) &\rightarrow \max \end{aligned}$$

Let us call it as the convex task.

Definition. The pair $\bar{x}^* \in \Pi(G_1), \bar{y}^* \in \Pi(G_2)$ such that $\max_{\bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2)} \varphi(\bar{x}, \bar{y}) - \varphi(\bar{x}^*, \bar{y}^*) \leq \varepsilon$ is called ε -solution of the convex task.

Theorem 9. For every ε the convex task can be ε -solved in polynomial time. The length of an input is a sum of the lengths of descriptions of $G_1 = (V, \succ^1)$, $G_2 = (V, \succ^2)$ and integer weights w_v . And obtained ε -solution (\bar{x}^*, \bar{y}^*) satisfies $|x_i^* - y_i^*| \leq \frac{1}{2}$.

Lemma. The pair $\overline{\xi}^{opt} = (\overline{x}^{opt}, \overline{y}^{opt}) = \arg \max_{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2)} \varphi(\overline{x}, \overline{y})$ satisfies $|x_i^{opt} - y_i^{opt}| \leq \frac{1}{2}$.

Proof of lemma. Quadratic functional $\varphi(\overline{x}, \overline{y})$ is not bounded in R^{2n} and its maximum on the set $\Pi(G_1) \times \Pi(G_2)$ is located on the borders of polyhedron. Let $\overline{a}_1^T \overline{\xi} \leq b_1, \dots, \overline{a}_s^T \overline{\xi} \leq b_s$ be those inequalities from the definition of polyhedron that turn into equalities. From the optimality of $(\overline{x}^{opt}, \overline{y}^{opt})$ it is clear that the cone $\{\overline{\xi} | \overline{a}_1^T \overline{\xi} \leq 0\} \cap \dots \cap \{\overline{\xi} | \overline{a}_s^T \overline{\xi} \leq 0\} \cap \{\overline{\xi} | \nabla_{\overline{\xi}^{opt}} \varphi(\overline{\xi}^{opt})^T \overline{\xi} > 0\} = \emptyset$. And then, from theorem of Farkas-Minkovski, we conclude that $\varphi(\overline{\xi}^{opt})$ can be expanded on positive combination of vectors $\overline{a}_1, \dots, \overline{a}_s$. But taking into account that components of those vectors are positive we obtain that $\nabla_{\overline{\xi}^{opt}} \varphi(\overline{\xi}^{opt}) = \|w_1(x_1^{opt} - y_1^{opt} + \frac{1}{2}), w_1(y_1^{opt} - x_1^{opt} + \frac{1}{2}), \dots, w_n(x_n^{opt} - y_n^{opt} + \frac{1}{2}), w_n(y_n^{opt} - x_n^{opt} + \frac{1}{2})\|^T \geq \overline{0}$. Lemma proved.

Proof of theorem. Since the function $\varphi(\overline{x}, \overline{y})$ is concave, the set of pairs

$$\begin{aligned} \overline{x} &\in \Pi(G_1) \\ \overline{y} &\in \Pi(G_2) \\ \varphi(\overline{x}, \overline{y}) &\geq c \\ -\frac{1}{2} &\leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n} \end{aligned}$$

is convex.

Note that for every given vector pair $\overline{x}', \overline{y}'$ the task of defining whether it belongs to the set $\Pi(G_1) \times \Pi(G_2)$ or not can be solved in polynomial time. Actually, by Floyd-Warshall algorithm we can find the longest path from s to t in orgraphs G_1 and G_2 in polynomial time, where by length of a path we mean a sum of weights of vertexes on the path. Comparing the results with 1 we see that if they are less than 1 then $\overline{x}', \overline{y}' \in \Pi(G_1) \times \Pi(G_2)$. Besides, if $\overline{x}', \overline{y}' \notin \Pi(G_1) \times \Pi(G_2)$ then the path which length is greater than 1 will give us a violated inequality in the definition of the polyhedron $\Pi(G_1) \times \Pi(G_2)$.

And for given $\overline{x}', \overline{y}'$, the satisfaction of conditions $\varphi(\overline{x}', \overline{y}') \geq c$, and in negative case, the separating hyperplane for the pair $\overline{x}', \overline{y}'$ and the set $\{(\overline{x}, \overline{y}) | \varphi(\overline{x}, \overline{y}) \geq c + \varepsilon\}$ can be found in polynomial time.

Actually,

$$\begin{aligned} \{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2) | (\nabla_{\overline{x}'} \varphi(\overline{x}', \overline{y}'), \overline{x} - \overline{x}') + (\nabla_{\overline{y}'} \varphi(\overline{x}', \overline{y}'), \overline{y} - \overline{y}') \geq \varepsilon\} &\supseteq \\ &\supseteq \{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2) | \varphi(\overline{x}, \overline{y}) \geq c + \varepsilon\} \end{aligned}$$

This can be seen from the following inequalities for concave quadratic function φ and points $(\overline{x}, \overline{y}), (\overline{x}', \overline{y}')$ such that $\varphi(\overline{x}, \overline{y}) \geq c + \varepsilon$ and $\varphi(\overline{x}', \overline{y}') \leq c$: $\varepsilon \leq \varphi(\overline{x}, \overline{y}) - \varphi(\overline{x}', \overline{y}') \leq (\nabla_{\overline{x}'} \varphi(\overline{x}', \overline{y}'), \overline{x} - \overline{x}') + (\nabla_{\overline{y}'} \varphi(\overline{x}', \overline{y}'), \overline{y} - \overline{y}')$.

Then rounding each component of vectors $\nabla_{\overline{x}'} \varphi(\overline{x}', \overline{y}')$ and $\nabla_{\overline{y}'} \varphi(\overline{x}', \overline{y}')$ to the first $2(\log n + |\log \varepsilon| + 1)$ symbols in binary representation and denoting them as c_x and c_y , will give us the separating hyperplane

$$\left\{ (\overline{x}, \overline{y}) | (c_x, \overline{x} - \overline{x}') + (c_y, \overline{y} - \overline{y}') \geq \frac{\varepsilon}{2} \right\}.$$

According to [3], in this case to find a pair \bar{x}', \bar{y}' that satisfies:

$$\begin{aligned}\bar{x}' &\in \Pi(G_1) \\ \bar{y}' &\in \Pi(G_2) \\ \varphi(\bar{x}', \bar{y}') &\geq c \\ -\frac{1}{2} &\leq x'_i - y'_i \leq \frac{1}{2}, i = \overline{1, n}\end{aligned}$$

can be done in polynomial time, or it will be shown that

$$\{(\bar{x}, \bar{y}) \mid \bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2), \varphi(\bar{x}, \bar{y}) \geq c + \varepsilon, -\frac{1}{2} \leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n}\} = \emptyset.$$

Taking into account that $|\varphi(\bar{x}, \bar{y})| \leq 2 \sum_{v \in V} w_v$, by the method of binary division we find such a constant c , that the set

$$\Omega = \{(\bar{x}, \bar{y}) \mid \bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2), \varphi(\bar{x}, \bar{y}) \geq c, -\frac{1}{2} \leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n}\} \neq \emptyset$$

and

$$\{(\bar{x}, \bar{y}) \mid \bar{x} \in \Pi(G_1), \bar{y} \in \Pi(G_2), \varphi(\bar{x}, \bar{y}) \geq c + \varepsilon, -\frac{1}{2} \leq x_i - y_i \leq \frac{1}{2}, i = \overline{1, n}\} = \emptyset.$$

From lemma we see that $\bar{\xi}^{opt} \in \Omega$ and $\varphi(\bar{\xi}^{opt}) < c + \varepsilon$. And every pair from Ω is an ε -solution of the task. Theorem proved.

Consider the following approximate algorithm for 2-MaxCMS.

1. Find a pair (\bar{x}', \bar{y}') such that $\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \varphi(\bar{x}, \bar{y}) \leq \varphi(\bar{x}', \bar{y}') + \varepsilon \quad |x'_i - y'_i| \leq \frac{1}{2}$

where $\varepsilon = \frac{1}{16}$.

2. Find $\bar{x}^* = \arg \max_{\bar{x} \in \Pi(G_1)} \psi(\bar{x}, \bar{y}')$ and $\bar{y}^* = \arg \max_{\bar{y} \in \Pi(G_2)} \psi(\bar{x}^*, \bar{y})$. There \bar{x}^*, \bar{y}^* are

boolean.

The answer of an algorithm is the set of vertexes $\{v \mid x_v^* y_v^* = 1\}$.

It is easy to see that all stages of the algorithm are polynomial. Let us investigate its answer.

Denote $W = \sum_{v \in V} w_v$ and $\varphi(\bar{x}', \bar{y}') = \alpha W$. It is clear that $0 \leq \alpha \leq 1$.

Theorem 10. The following is true

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \left(\frac{1}{4} - \left(\alpha - \frac{1}{2} \right)^2 \right) W + \varepsilon,$$

if $\alpha \geq \frac{1}{2}$. And also

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \frac{1}{4} W + \varepsilon,$$

when $\frac{3}{8} \leq \alpha \leq \frac{1}{2}$. And

$$\max_{(\bar{x}, \bar{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\bar{x}, \bar{y}) - \psi(\bar{x}^*, \bar{y}^*) \leq \left(\frac{1}{4} - \left(\alpha - \frac{3}{8} \right)^2 \right) W + \varepsilon,$$

if $\alpha \leq \frac{3}{8}$.

Proof. Let us bound $\varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}')$, using the fact of concavity of $f(x) = x - x^2$:

$$\begin{aligned}\varphi(\bar{x}', \bar{y}') - \psi(\bar{x}', \bar{y}') &= \sum_{v \in V} \frac{1}{2} w_v (x'_v - x_v'^2) + \frac{1}{2} w_v (y'_v - y_v'^2) \leq \\ &\leq \sum_{v \in V} w_v \frac{(x'_v + y'_v)}{2} \left(1 - \frac{(x'_v + y'_v)}{2} \right) = \sum_{v \in V} w_v \left(\frac{1}{4} - \left(\frac{x'_v + y'_v - 1}{2} \right)^2 \right)\end{aligned}$$

When $\alpha \geq \frac{1}{2}$:

$$\alpha W = \varphi(\overline{x'}, \overline{y'}) = \sum_{v \in V} -\frac{1}{2}w_v(x'_v - y'_v)^2 + \frac{1}{2}w_v y'_v + \frac{1}{2}w_v x'_v \leq \sum_{v \in V} \frac{1}{2}w_v y'_v + \frac{1}{2}w_v x'_v$$

and from this:

$$\sum_{v \in V} w_v \frac{(x'_v + y'_v - 1)}{2} \geq \left(\alpha - \frac{1}{2}\right) W.$$

Then

$$\varphi(\overline{x'}, \overline{y'}) - \psi(\overline{x'}, \overline{y'}) \leq \sum_{v \in V} w_v \left(\frac{1}{4} - \left(\frac{x'_v + y'_v - 1}{2} \right)^2 \right) \leq \frac{1}{4}W - t,$$

where $t = \min_{\sum_{v \in V} w_v t_v \geq (\alpha - \frac{1}{2})W} \sum_{v \in V} w_v t_v^2$. It is obvious that $t = \left(\alpha - \frac{1}{2}\right)^2 W$. So, we obtain

$$\varphi(\overline{x'}, \overline{y'}) - \psi(\overline{x'}, \overline{y'}) \leq \left(\frac{1}{4} - \left(\alpha - \frac{1}{2} \right)^2 \right) W.$$

Then using $\varphi(\overline{x'}, \overline{y'}) \geq \max_{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\overline{x}, \overline{y}) - \varepsilon$ and $\psi(\overline{x^*}, \overline{y^*}) \geq \psi(\overline{x'}, \overline{y'})$ we finally obtain:

$$\max_{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\overline{x}, \overline{y}) - \psi(\overline{x^*}, \overline{y^*}) \leq \left(\frac{1}{4} - \left(\alpha - \frac{1}{2} \right)^2 \right) W + \varepsilon.$$

Almost analogous, when $\alpha \leq \frac{3}{8}$,

$$\begin{aligned} \alpha W &= \varphi(\overline{x'}, \overline{y'}) = \sum_{v \in V} -\frac{1}{2}w_v(x'_v - y'_v)^2 + \frac{1}{2}w_v y'_v + \frac{1}{2}w_v x'_v \geq \\ &\geq \sum_{v \in V} \frac{1}{2}w_v y'_v + \frac{1}{2}w_v x'_v - \frac{1}{8}W \end{aligned}$$

and from this:

$$\sum_{v \in V} w_v \frac{(x'_v + y'_v - 1)}{2} \leq \left(\alpha - \frac{3}{8} \right) W.$$

Analogously,

$$\varphi(\overline{x'}, \overline{y'}) - \psi(\overline{x'}, \overline{y'}) \leq \frac{1}{4}W - s,$$

where $s = \min_{\sum_{v \in V} w_v t_v \leq (\alpha - \frac{3}{8})W} \sum_{v \in V} w_v t_v^2 = \left(\alpha - \frac{3}{8} \right)^2 W$. And finally,

$$\begin{aligned} \max_{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\overline{x}, \overline{y}) - \psi(\overline{x^*}, \overline{y^*}) &\leq \varphi(\overline{x'}, \overline{y'}) - \psi(\overline{x'}, \overline{y'}) + \varepsilon \leq \\ &\leq \left(\frac{1}{4} - \left(\alpha - \frac{3}{8} \right)^2 \right) W + \varepsilon. \end{aligned}$$

The statement of the theorem in case when $\frac{3}{8} \leq \alpha \leq \frac{1}{2}$ is obvious. The theorem proved.

6 Conclusion

As mentioned above, MaxCMS can be considered as a subcase for the vertex cover problem. From this point of view the task of finding MaxCMS is equivalent to the task of removing "noisy" objects from the training set with a minimal total weight. Let us compare the approximation ratio of our algorithm with a well-known, standard 2-approximation of vertex cover, that can be found for any graph with weighted vertexes in polynomial time[1].

It is clear that ε can be made arbitrarily small and it does not play any role in the bound of theorem 10 because the bounded value is integer. So, for simplicity, we will believe that $\varepsilon = 0$. Let us denote $\varphi(\overline{x'}, \overline{y'}) = \alpha W \geq W - \Delta = \max_{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\overline{x}, \overline{y})$.

It is obvious that the ratio 2 of approximation has a meaning only if maximal consistent with monotonicity set of a special orgraph has a weight more than half of the sum of weights of all vertexes, i.e. $\alpha \geq \alpha' = \frac{MaxCMS}{W} \geq \frac{1}{2}$. In this case from theorem 10 we obtain that:

$$\max_{(\overline{x}, \overline{y}) \in \Pi(G_1) \times \Pi(G_2)} \psi(\overline{x}, \overline{y}) - \psi(\overline{x}^*, \overline{y}^*) \leq \alpha(1 - \alpha)W \leq \alpha'(1 - \alpha')W = \alpha'\Delta$$

which means that our algorithm has an approximation ratio equal to $1 + \alpha' \leq 2$.

For "almost correct" data, i.e. when $\alpha' \approx 1$, algorithm has an approximation ratio close to standard 2. But for "noisy" data it appears to be better than standard. For extreme case when $\alpha' \approx \frac{1}{2}$ standard 2-approximation means there is no guarantee that we will not remove all objects as "noise". On the contrary, the total weight of objects removed by our algorithm in any case can not exceed optimal solution by more than $\frac{1}{4}W$. And the bound of theorem 10 shows that our algorithm can find good approximations to MaxCMS for cases when even more than half of the training set consists of "noisy" data ($\alpha \leq \frac{3}{8}$).

References

- [1] D. S. Hochbaum, Approximation algorithms for the set covering and vertex cover problems. SIAM Journal on Computing, 11:555–556, 1982.
- [2] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman, 1979.
- [3] Grotshel M., Lovasz L., Schrijver A, Geometric algorithms and combinatorial optimization. Springer-Verlag, Berlin Geidelberg New York, 1988.
- [4] Mohring R.H. Algorithmic aspects of comparability graphs and interval graphs. In Graphs and Order, pp.41-101. Dordrecht: Reidel, 1985.